Report of Data Mining HW1

109550027 紀竺均

1. How do you select features for your model input, and what preprocessing did you perform?

a. Data Preprocessing: First, I loaded the csv file into panda dataframe, then I selected the data of 0am-9am for every element. Next, I replace non-numeric values with NaN values, and fill NaN values with the median of their respective columns. Finally, I reshape the data such that each element becomes a separate feature, which means that each row represents a day, but each column represents an hour-element combination.

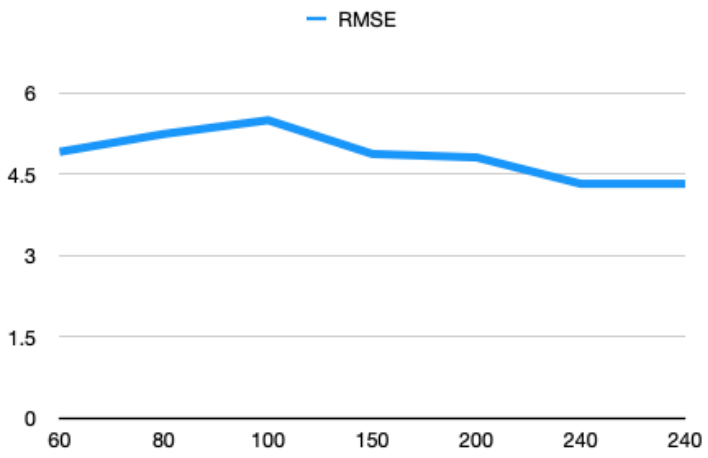
b. Feature Selection: I correlate all the element's nine-hour-mean with targeted PM2.5 value (that is, the value of PM2.5 at 10am), and select the features with the top five highest correlation.(Five being the most efficient constant after multiple attempts.) This element-feature selection is also being applied on testing data.

2. Compare the impact of different amounts of training data on the PM2.5 prediction accuracy. Visualize the results and explain them.

I calculate the RMSE loss of different amounts of training data. Thanks to the preprocessing I have done, I use the number of days on behalf of the amount of training data. For simplicity, I trained for 50000 iterations only.

Initially, with small amounts of data, the model performs poorly due to insufficient data (or overfitting). As more data is added, the RMSE loss decreases significantly.

#days	RMSE
60	4.909
80	5.238
100	5.492
150	4.87
200	4.806
240	4.32



3. Discuss the impact of regularization on PM2.5 prediction accuracy.

After I applied regularization on my model, my RMSE loss became smaller. I think there are two main reasons:

a. Prevent Overfitting. Regularization penalizes the magnitude of the coefficients in linear regression models, thus helping to reduce overfitting and gain better result scores in testing data's loss.
b. Handling Multicollinearity. Regularization is particularly useful in cases where features are highly correlated, in my case, I have done feature selection and extracted the highly correlated data, thus, regularization can make the linear model more sensitive to minor changes in the input.